Quiz

1. **Which of the following is a benefit of using denormalized data structures in BigQuery?**
   1. Reduces the amount of data processed
   2. *Increases query speed*
   3. Reduces the amount of storage required
   4. Provides internal referential integrity for inner joins
2. **Which of these sources can you NOT load data into BigQuery from?**
   1. File upload
   2. Google Drive
   3. Google Cloud Storage
   4. *Managed Cloud SQL instance.*
3. **Which of the following pairs of methods that can be used to denormalize tables in BigQuery?**
   1. Split table into multiple tables and use a partitioned table
   2. *Join tables into one table and use nested repeated fields*
   3. Use a partitioned table and join table into one tables
   4. Use nested repeated fields and use table column clustering
4. **If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?**
   1. Unsupervised learning
   2. *Regressor*
   3. Classifier
   4. Clustering estimator
5. **Does Dataflow process batch data pipelines or streaming data pipelines?**
   1. Only Batch Data Pipelines
   2. *Both Batch and Streaming Data Pipelines*
   3. Only Streaming Data Pipelines
   4. None of the above
6. **Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?**
   1. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
   2. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
   3. *Create a Cloud Dataproc cluster that uses the Google Cloud Storage.*
   4. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.
7. **Your company’s on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?**
   1. *Put the data into Google Cloud Storage.*
   2. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
   3. Put the data into a BigQuery dataset.
   4. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Cloud SQL.
8. **You are using Cloud Pub/Sub to stream inventory updates from many point-of-sale (POS) terminals into BigQuery. Each update event has the following information:   
    product identifier “prodSku”,   
    change increment “quantityDelta”,   
    POS identification “termId”,   
    “messageId” which is created for each push attempt from the terminal.   
   During a network outage, you discovered that duplicated messages were sent, causing the inventory system to over-count the changes. You determine that the terminal application has design problems and may send the same event more than once during push retries. You want to ensure that the inventory update is accurate. What should you do?**
   1. Inspect the “publishTime” of each message. Make sure that messages whose “publishTime” values match rows in the BigQuery table are discarded.
   2. Inspect the “messageId” of each message. Make sure that any messages whose “messageId” values match corresponding rows in the BigQuery table are discarded.
   3. Instead of specifying a change increment for “quantityDelta”, always use the derived inventory value after the increment has been applied. Name the new attribute “adjustedQuantity”.
   4. *Add another attribute orderId to the message payload to mark the unique check-out order across all terminals. Make sure that messages whose “orderId” and “prodSku” values match corresponding rows in the BigQuery table are discarded.*
9. **Why do you need to split a machine learning dataset into training data and test data?**
   1. So you can try two different sets of features
   2. *To make sure your model is generalized for more than just the training data*
   3. To allow you to create unit tests in your code
   4. So you can use one dataset for a wide model and one for a deep model
10. **The Dataflow product is a runner for job developed using which Apache product?**
    1. Apache Spark
    2. Apache Hadoop
    3. Apache Kafka
    4. *Apache Beam*
11. **Cloud Composer runs instances of which Apache product?**
    1. Beam
    2. Calcite
    3. *Airflow*
    4. Fusion
12. **Cloud Dataproc is a managed Apache Hadoop and Apache \_\_\_\_\_ service.**
    1. Blaze
    2. *Spark*
    3. Fire
    4. Ignite
13. **Which of the following is NOT true about Dataflow pipelines?**
    1. Pipelines are a set of operations
    2. Pipelines represent a data processing job
    3. Pipelines represent a directed graph of steps
    4. *Pipelines can share data between instances*
14. **What data product would you use for transactional data that will exceed the recommended size of a managed Cloud SQL database?**
    1. BigQuery
    2. Firestore
    3. *Cloud Spanner*
    4. Big Table
15. **Cloud Bigtable is Google’s \_\_\_\_\_\_ Big Data database service.**
    1. Relational
    2. mySQL
    3. *NoSQL*
    4. SQL Server